# ToTeM: une méthode de détection de communautés adaptées aux réseaux d'information

David Combe*, Christine Largeron*
Előd Egyed-Zsigmond**, Mathias Géry*

*Université de Lyon, F-42023, Saint-Étienne, France,
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
{david.combe, christine.largeron, mathias.gery}@univ-st-etienne.fr
**Université de Lyon
UMR 5205 CNRS, LIRIS
7 av J. Capelle, F 69100 Villeurbanne, France
elod.egyed-zsigmond@insa-lyon.fr

**Résumé.** Alors que les réseaux sociaux s'attachaient à représenter des entités et les relations qui existaient entre elles, les réseaux d'information intègrent également des attributs décrivant ces entités ; ce qui conduit à revisiter les méthodes d'analyse et de fouille de ces réseaux. Dans cet article, nous proposons une méthode de classification des sommets d'un graphe qui exploite d'une part leurs relations et d'autre part les attributs les caractérisant. Cette méthode reprend le principe de la méthode de Louvain en l'étendant de façon à permettre la manipulation d'attributs continus d'une manière symétrique à ce qui existe pour les relations.

## 1 Introduction

L'objectif de la détection de communautés dans les graphes, ou encore dans les réseaux sociaux, est de créer une partition des sommets, en tenant compte des relations qui existent entre ces sommets dans le graphe, de telle sorte que les communautés soient composées de sommets fortement connectés (Fortunato (2010)). Ainsi, les principales méthodes de détection de communautés proposées dans la littérature se concentrent sur la structure des liens, en ignorant les propriétés des sommets. Or dans de nombreuses applications, les réseaux sociaux peuvent être représentés par des graphes dont les sommets ont des attributs qui peuvent être pris en compte pour détecter plus efficacement les communautés. Ceci a conduit à revisiter cette problématique afin d'opérer cette détection non seulement à partir des relations décrites par le graphe, mais aussi à partir d'attributs caractérisants les sommets et cela a donné lieu récemment à l'introduction de méthodes qui exploitent ces deux types de données (Moser et al. (2007); Zhou et al. (2009); Li et al. (2008); Cruz Gomez et al. (2011); Combe et al. (2012); Dang et Viennet (2012)).

Dans cet article, nous proposons ToTeM, une méthode de classification de graphes à vecteurs d'attributs qui reprend le principe de la méthode de Louvain, basée sur l'optimisation

de la modularité, en l'étendant de façon à permettre la prise en compte d'attributs numériques d'une manière symétrique à ce qui existe pour les relations (Blondel et al. (2008)). Après avoir défini plus formellement le problème de la détection de communautés dans un réseau d'information dans la section 2, nous rappelons brièvement le principe de la méthode de Louvain et introduisons ToTeM dans la section 3 avant de décrire des critères globaux de partitionnement dans la section suivante.

## 2 Énoncé du problème et notations

Étant donné un graphe $G = (V, E)$ où $V = \{v_1, \ldots, v_i, \ldots, v_n\}$ est l'ensemble des sommets et $E \subset V \times V$ est l'ensemble des arêtes non étiquetées. On suppose que chaque sommet $v_i \in V$ est associé à un vecteur $d_i = (w_{i1}, \ldots, w_{ij}, \ldots, w_{iT})$ à valeurs réelles de sorte que $G$ forme un réseau d'information (Zhou et al. (2009)). Dans un problème de partitionnement de réseau d'information, les liens et les attributs sont considérés, de telle sorte que d'une part il doit y avoir de nombreuses arêtes au sein de chaque classe et relativement peu entre elles et d'autre part, deux sommets appartenant à la même classe sont plus proches en termes d'attributs que deux sommets appartenant à des classes différentes. Ainsi, l'objectif est de partitionner l'ensemble $V$ des sommets en $r$ classes disjointes formant une partition $\mathscr{P} = \{C_1, \ldots, C_r\}$ où $r$ est a priori inconnu et de telle sorte que les sommets appartenant à un même groupe soient connectés et homogènes vis-à-vis des attributs. Dans la suite, on notera $A$ la matrice d'adjacence de $G$ telle que $A_{ij}$ indique la valuation de l'arête entre $i$ et $j$ si elle existe et vaut 0 s'il n'existe pas d'arête entre $i$ et $j$. Le degré du sommet $i$, noté $k_i$, est égal à $\sum_j A_{ij}$ et $c_i$ désignera la classe d'appartenance de $i$ dans la partition $\mathscr{P}$.

## 3 La méthode ToTeM

La méthode ToTeM que nous proposons est une extension de la méthode de Louvain qui consiste elle-même à optimiser le critère de modularité (Blondel et al. (2008); Newman et Girvan (2004)) :

$$Q(\mathscr{P}) = \frac{1}{2m}\Sigma_{ii'}\left[(A_{ii'} - \frac{k_i \cdot k_{i'}}{2m}) \cdot \delta(c_i, c_{i'})\right] \quad (1)$$

où $(i, i')$ prend toutes les valeurs de $V \times V$, $m$ est la somme des poids de toutes les arêtes du graphe et $\delta$ est la fonction de Kronecker qui vaut 1 si ses arguments sont égaux et 0 sinon.

L'algorithme comporte deux phases. À partir de la partition discrète, la première phase consiste à essayer de déplacer successivement chaque sommet vers la classe de ses voisins et à l'affecter à la classe ayant apporté le plus fort gain de modularité. Lorsque plus aucune amélioration n'est possible, dans une seconde phase, un nouveau graphe pondéré est formé à partir des classes obtenues à l'issue de la première phase. Chaque classe devient un sommet du nouveau graphe et une arête entre deux sommets a pour poids la somme des poids des arêtes présentes entre des sommets contenus précédemment dans les classes correspondantes. Les deux phases sont répétées jusqu'à ce qu'il n'y ait plus de modification possible. Le gain de

modularité induit par le déplacement d'un sommet isolé $x$ vers une classe $C_l$ est égal à :

$$\Delta Q = \left[\frac{\sum_{in}+2k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad (2)$$

où $\sum_{in}$ est la somme des poids des arêtes ayant leurs deux extrémités dans la classe $C_l$, $\sum_{tot}$ est la somme des poids des arêtes adjacentes aux sommets de $C_l$, $k_{i,in}$ est la somme des poids des arêtes de $i$ aux sommets de $C_l$ (Blondel et al. (2008)).

La méthode ToTeM, que nous introduisons dans l'Algorithme 1, repose sur l'optimisation d'un critère global permettant de classer les sommets en se souciant à la fois de la qualité des classes d'un point de vue relationnel mais également du point de vue des attributs.

**Algorithme 1** : ToTeM

**Entrées** : Réseau d'information $G$
**Sorties** : Partition $\mathscr{P}$

```
répéter
    Placer chaque sommet de G dans une unique classe;
    Sauver la qualité globale de cette décomposition;
    tant que il y a des sommets déplacés faire
        pour tous les sommet x de G faire
            l ← classe voisine maximisant le gain du critère d'évaluation;
            si l induit un gain strictement positif alors
                déplacer x de sa classe vers l;
    si le critère de qualité atteint est supérieur à sa valeur initiale alors
        fin ← faux;
        Afficher la décomposition trouvée;
        Transformer G en le réseau d'information entre les classes;
    sinon
        fin ← vrai ;
jusqu'à fin ;
```

Pour ce qui est de la qualité par rapport aux relations, on peut retenir la modularité. Le gain est alors mesuré suivant la formule 2. Pour ce qui est de la qualité par rapport aux attributs, plusieurs mesures sont envisageables comme le taux d'inertie inter-classes ou l'indice de Calinski, détaillés dans les sections suivantes.

# 4 Gain d'inertie et critères de qualité globale

Il est possible d'améliorer l'efficacité de l'algorithme ToTeM lorsque le critère global est basé sur l'inertie inter-classes d'une partition en remarquant que la variation d'inertie inter-classes induite par la réaffectation d'un sommet peut être calculée uniquement à l'aide d'information locale. Etant donné $V$ l'ensemble des $n$ sommets du graphe représentés dans un

espace vectoriel défini par les attributs et muni d'une distance euclidienne à laquelle est associée une norme $\|\cdot\|$. A tout sommet $x$ de $V$ est associé un poids positif $m_x$ et, sans perte de généralité, on peut supposer qu'initialement tous les sommets ont le même poids. On note $g$ le centre de gravité de $V$ et pour toute classe $C_l$ de $\mathscr{P}$, $g_l$ son centre de gravité et $m_l$ la somme des poids des éléments de cette classe $C_l$. Considérons deux partitions $\mathscr{P}$ et $\mathscr{P}'$ telles que $\mathscr{P} = (A, B, C_1, \ldots, C_r)$ et $\mathscr{P}' = (A \setminus \{x\}, B \cup \{x\}, C_1, \ldots, C_r)$. Par la suite, $A \setminus \{x\}$ désigne la classe $A$ privée du sommet $x$ et $B \cup \{x\}$ la classe $B$ augmentée du sommet $x$.

L'inertie inter-classes $I_{inter}(\mathscr{P})$ associée à la partition $\mathscr{P}$ est égale à :

$$I_{inter}(\mathscr{P}) = m_A \|g_A - g\|^2 + m_B \|g_B - g\|^2 + \sum_{l=1,..,r} m_l \|g_l - g\|^2 \tag{3}$$

où $g_A$ (respectivement $g_B$) est le centre de gravité de $A$ (respectivement $B$) et $m_A$ (respectivement $m_B$) est le poids de $A$ (respectivement $B$). L'inertie inter-classes de la partition $\mathscr{P}'$ obtenue en retirant $x$ de sa classe $A$ et en l'affectant à la classe $B$ vaut :

$$I_{inter}(\mathscr{P}') = m_{A\setminus\{x\}} \left\|g_{A\setminus\{x\}} - g\right\|^2 + m_{B\cup\{x\}} \left\|g_{B\cup\{x\}} - g\right\|^2 + \sum_{l=1,..,r} m_l \|g_l - g\|^2 \tag{4}$$

La variation d'inertie inter-classes induite par le déplacement du sommet $x$ de la classe $A$ vers la classe $B$ est donnée par :

$$\begin{aligned} \Delta I_{inter} =& (m_A - m_x) \cdot \left\|g_{A\setminus\{x\}} - g\right\|^2 + (m_B + m_x) \cdot \left\|g_{B\cup\{x\}} - g\right\|^2 \\ & - m_A \cdot \|g_A - g\|^2 - m_B \cdot \|g_B - g\|^2 \end{aligned} \tag{5}$$

$g_{A\setminus\{x\}}$ et $g_{B\cup\{x\}}$ sont eux aussi calculés facilement en utilisant seulement l'effectif représenté par le sommet $x$ et les classes $A$ et $B$ ainsi que leurs centres de gravités $g_A$, $g_B$ :

$$g_{A\setminus\{x\}} = \frac{1}{m_{A-\{x\}}} \sum_{i\in A\setminus\{x\}} m_i d_i = \frac{1}{m_A - m_x}(m_A \cdot g_A - m_x \cdot d_x) \tag{6}$$

$$g_{B\cup\{x\}} = \frac{1}{m_B + m_x}(m_B \cdot g_B + m_x \cdot d_x) \tag{7}$$

Les valeurs des poids associés aux classes peuvent aussi être recalculées à l'aide de l'information locale : $m_{A\setminus\{x\}} = m_A - m_x$ et $m_{B\cup\{x\}} = m_B + m_x$

## 4.1 Synthèse des informations du graphe et des attributs dans la seconde phase

L'opération de synthèse des informations du graphe consiste, à l'instar de ce qui est opéré dans la méthode de Louvain, à fusionner les sommets affectés à une même classe de façon à n'en faire qu'un seul sommet. Ainsi, à partir de la partition $\mathscr{P}' = (C_1, ..., C_r)$ obtenue à l'issue de la première phase, un nouveau graphe $G' = (V', E')$ est crée. Ce graphe comporte autant de sommets qu'il y a de classes dans $\mathscr{P}'$ et chaque sommet $v'_l$ de $V'$ incarne une classe $C_l$ de $\mathscr{P}'$. La valuation de l'arête éventuellement présente entre les sommets $v'_y$ et $v'_z$ de $V'$ est égale à la somme des valuations des arêtes présentes entre des sommets de $G$ appartenant aux

classes $C_y$ et $C_z$ de $\mathscr{P}'$ qui ont été représentées par $v'_y$ et $v'_z$ dans $V'$. Soit $\tau$ la fonction qui indique, pour un sommet de $V$, par quel sommet de $V'$ il est représenté, alors le poids associé à une arête se calcule de la façon suivante :

$$weight(v'_y, v'_z) = \sum_{(v_a, v_b) \in V \times V} A(v_a, v_b) \cdot \delta(\tau(v_a), v'_y) \cdot \delta(\tau(v_b), v'_z) \tag{8}$$

Enfin, les arêtes internes aux classes de $P$ deviennent des boucles dans $G'$.

De plus, il est nécessaire de transférer les informations relatives aux attributs sur le nouveau graphe $G'$. Pour cela, on affecte les poids des classes d'origine aux sommets de destination et le centre de gravité de la classe d'origine devient le vecteur d'attributs du sommet de destination. Ainsi, pour tout sommet $v'_l$ de $V'$ résultant de la classe $C_l$ de $\mathscr{P}'$ on a $m_{v'_l} = m_{C_l}$ et $d_{v'_l} = g_{C_l}$

## 4.2 Critères de qualité globale

Le critère de qualité globale intervenant dans l'algorithme ToTeM doit être une fonction d'une mesure de qualité de la partition par rapport aux relations et d'une mesure de sa qualité par rapport aux attributs. La modularité peut être utilisée comme mesure de la qualité par rapport aux relations. Pour ce qui est de la qualité par rapport aux attributs, une première solution envisageable peut consister à prendre le taux d'inertie inter-classes. Ce qui conduit à une première mesure de qualité globale définie par :

$$CG1 = \frac{I_{inter}(\mathscr{P})}{I(\mathscr{P})} \cdot mod(\mathscr{P}) \tag{9}$$

où $I$ désigne la variance totale de $V$. Cependant, le taux d'inertie inter-classes n'est pas conçu pour comparer des partitions ayant un nombre de classes différent. En effet, il varie structurellement avec le nombre de classes de la partition de sorte qu'il est maximum pour la partition discrète. Une solution simple visant à palier ce biais structurel consiste à tenir compte du nombre de classes de la partition pour définir un critère global :

$$CG2 = \frac{I_{inter}(\mathscr{P})}{|\mathscr{P}| \cdot I(\mathscr{P})} \cdot mod(\mathscr{P}) \tag{10}$$

où $|\mathscr{P}|$ est le nombre de classes de la partition $\mathscr{P}$.

Contrairement au précédent, ce critère donne un avantage aux partitions à faible nombre de classes. Une alternative pour palier cet inconvénient consiste à avoir recours à des indices conçus dans le but de déterminer le nombre de classes de le cas du partitionnement de données vectorielles, comme par exemple l'indice de Calinski-Harabasz, celui de Dunn ou celui de Davies-Bouldin (Calinski et Harabasz (1974); Davies et Bouldin (1979)).

Une autre solution pour comparer deux partitions $\mathscr{P}$ et $\mathscr{P}'$ de taille respective $r$ et $r'$ consiste à utiliser la probabilité critique résultant de tests de comparaison de variance. En effet, sous l'hypothèse nulle selon laquelle les classes ne sont pas significativement différentes au sein de la partition $\mathscr{P}$, la statistique $F(\mathscr{P})$ définie par :

$$F_{\mathscr{P}} = \frac{V_{inter}/(r-1)}{(V_T - V_{inter})/(n-r))} \tag{11}$$

suit une loi de Fisher-Snedecor $F(r-1, n-r)$ à $(r-1, n-r)$ degrés de liberté où $V_{inter}$ désigne la variance entre les classes et $V_T$ la variance totale. On peut donc calculer la probabilité critique $PC$ associée : $PC = P(F(r-1, n-r) > F_{\mathscr{P}})$. De même, la statistique $F(\mathscr{P}')$ peut être déterminée sur la partition $\mathscr{P}'$ et on peut en déduire $PC'$. La comparaison des probabilités critiques associées à $\mathscr{P}$ et à $\mathscr{P}'$ conduira à préférer la partition pour laquelle cette probabilité est la plus faible.

**Remerciements** Ce travail est partiellement soutenu par la région Rhône-Alpes et St-Etienne Metropole (http://www.agglo-st-etienne.fr/)

## Références

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics P10008*.

Calinski, T. et J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics Theory and Methods 3*(1), 1–27.

Combe, D., C. Largeron, E. Egyed-Zsigmond, et M. Géry (2012). Getting clusters from structure data and attribute data. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM*, pp. 731–733.

Cruz Gomez, J. D., C. Bothorel, et F. Poulet (2011). Entropy based community detection in augmented social networks. In *Computational Aspects of Social Networks (CASoN)*, pp. 163–168.

Dang, T. A. et E. Viennet (2012). Community Detection based on Structural and Attribute Similarities. In *International Conference on Digital Society (ICDS)*, pp. 7–12.

Davies, D. et D. Bouldin (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence* (2), 224—-227.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*(3-5), 75–174.

Li, H., Z. Nie, W.-C. W. Lee, C. L. Giles, et J.-R. Wen (2008). Scalable Community Discovery on Textual Data with Relations. In *CIKM*, pp. 1203–1212.

Moser, F., R. Ge, et M. Ester (2007). Joint Cluster Analysis of Attribute and Relationship Data Without A-Priori Specification of the Number of Clusters. In *SIGKDD*, pp. 510.

Newman, M. et M. Girvan (2004). Finding and evaluating community structure in networks. *Physical review E 69*(2), 1–16.

Zhou, Y., H. Cheng, et J. Yu (2009). Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment 2*(1), 718–729.

## Summary

While social networks used to be handled through relations only, information networks are integrating attributes describing the entities. This lead us to revisit analysis and mining methods for these networks. We present a combined classification method of the vertices of a graph taking into account relations and attributes. This method uses the principles of the Louvain method extending it allowing the attribute manipulation in a symmetrical manner to what has been done for relations.